Document Title:	How It Got There: Associating Individual DNA Profiles with Specific Body Fluids in Mixtures Using Targeted Digital Gene Expression and RNA-SNP Identification
Author(s):	Jack Ballantyne, Ph.D., Erin Hanson
Document Number:	253930
Date Received:	October 2019
Award Number:	2014-DN-BX-K019

How it got there: associating individual DNA profiles with specific body fluids in mixtures using targeted digital gene expression and RNA-SNP identification

(Grant No. 2014-DN-BX-K019)
Project State Date: 1 January 2015
Project End Date: 31 December 2018

Final Summary Overview
Final report: Yes

Submitted: September 28, 2018

Prepared and submitted by:

Jack Ballantyne, Ph.D. (Principal Investigator)
Professor, Department of Chemistry
Associate Director (Research), National Center for Forensic Science
University of Central Florida
P.O. Box 162367
Orlando, FL 32816-2367

Co-authors: Erin Hanson (UCF, Co-PI)

Recipient organization:
University of Central Florida
4000 Central Florida Boulevard
Orlando, FL 32816

Prepared for:
National Institute of Justice
810 7th Street, NW
Washington, DC 20531
Program manager: Theodore Robinson

Table of Contents

I. Purpose

The facts that are at issue during a medical-legal investigation include facts that pertain to who was involved or what happened or why, how, where or when did an incident happen. The forensic genetic revolution enables one to be able to routinely answer the question as to who was involved (via DNA profiling). This is equivalent to source attribution in Cook and Evett's 'hierarchy of propositions' case interpretation paradigm. The physical 'activity' that led to the deposition of the DNA source is the next level in the case hierarchy and this is comprised of the what, how, where or when subset of the facts that are at issue. This work sought to address this activity level, namely how did the DNA profile get there (i.e. deposited on a particular substrate)? As an example, and depending upon the context of a case, a victim's DNA profile originating from her saliva versus originating from her vaginal secretions could have drastic implications with respect to the investigation and prosecution of the case. The ability to definitively identify the body fluid or tissue source of origin of DNA profiles in single or admixed samples would therefore provide some indication of the activity preceding the deposition of the DNA profile. In order for this to be realized in casework operations, there are two problems that need to be addressed. Firstly, we need an established and validated molecular based method for body fluid identification that is definitive and is capable of identifying, and distinguishing between, all of the commonly encountered body fluids and tissues, defined here as blood, semen, saliva, vaginal secretions, menstrual blood and skin. Traditional biochemical methods do not exist for all of these body fluids and tissues. Despite some significant advances in epigenetics and proteomics there exists at the present time only one fully validated molecular based method for comprehensive body fluid identification, namely mRNA profiling. However, primarily due to the unavailability of commercial kits, it hasn't become established yet in the US. This project sought to develop a next generation RNA-sequencing system for body fluid identification that will be platform-compatible with the forensic genomics NGS DNA assays under development and that are, arguably, likely to supplant standard CE based assays.

The second problem is that despite the ability to definitively identify the body fluids present in a mixture, it is not possible to associate the component DNA profiles with specific body fluids, a requirement in order to meet the goal of obtaining probative objective 'activity level' information in criminal investigations. Coding region SNPs (or RNA-SNPs), judiciously chosen to be present in the body fluid specific mRNA biomarkers targeted and sequenced as part of the body fluid identification assay will, for the first time, permit an association of a DNA profile with a specific body fluid or tissue in admixed samples. It is important to note that this DNA profile-body fluid association cannot be obtained using any other body fluid identification method (e.g. CE or real-time PCR analysis, epigenetics, proteomics).

The purpose of this project was to provide a novel piece of investigative information that to date has not been possible with other body fluid identification methods or analytical platforms, but could also result in a fully validated commercial body fluid identification product for immediate and facile transfer to those operational crime laboratories in the process of implementing NGS-based DNA analysis.

II. Project design

The purpose of this project was to utilize multiplexed targeted re-sequencing of mRNA transcripts in order to identify the body fluid or tissue of origin and to uniquely associate a specific DNA profile (i.e. donor of the stain) with the specific body fluid type, the latter to be accomplished through the analysis of coding region SNPs (cSNPs, which we also refer to as RNA-SNPs) within each individual mRNA target.

The goals were to: 1) investigate the use of next generation sequencing (NGS) for the definitive identification of forensically relevant biological fluids and tissues (blood, semen, saliva, vaginal secretions, menstrual blood and skin) and the use of simultaneously obtained coding region SNP genotypes to identify the donor of each body fluid or tissue in admixed biological stains; 2) develop a highly multiplexed panel that included the analysis of numerous gene targets for each fluid or tissue for the simultaneous recovery of digital gene expression data (body fluid identification) and RNA-SNP genotyping data (association of a specific DNA profile to an individual body fluid or tissue); 3) test and optimize the body fluid assay using forensic type samples (including studies on sensitivity, specificity, mixtures and performance with *bona fide* non-probative casework samples); 4) develop appropriate metrics and statistical methods for data interpretation; and 5) provide the developed assay to operational crime laboratories who could perform a casework evaluation and validation, including blind testing, to further test the accuracy and reliability of the assay.

III. Methods

**detailed protocols are provided in the 5 publications (Appendix A) resulting from this work*

<u>Multiplexed targeted mRNA NGS assay for body fluid identification - Illumina MiSeq platform</u>

NGS libraries of targeted body fluid gene candidates were prepared using the TruSeq® Targeted RNA kit (January 2016 protocol version; Illumina Inc., San Diego, CA). Total RNA input ranged from 50 – 100 ng. Pooled libraries were quantitated using the 2200 TapeStation (Agilent Technologies, Santa Clara, CA) and High Sensitivity D1000 Screen Tape according to the manufacturer's protocol. Pooled libraries were diluted and denatured according to the manufacturer's recommended protocol and a 600 µl 6 pM sample was pipetted into the MiSeq® V3 150 cycle reagent cartridge for sequencing on the MiSeq instrument. Sequencing was performed using 51 cycles (single read). Local sequencing software on the MiSeq analyzed the data (base calling, demultiplexing and alignment to the provided manifest file using a banded Smith Waterman alignment). A minimum sample total read count (MTR) of 5000 was used (samples below this threshold were excluded). A minimum biomarker read count (MBR) of 500 was used as an individual biomarker threshold (counts per biomarker below 500 were removed). A third threshold was then used in which individual biomarker read count values that were less than 0.5% of the total reads for the sample were also removed. Bar graphs of threshold-filtered counts were prepared by sample and by gene to evaluate gene expression and specificity. The percent contribution of reads was next calculated in order to provide the percentage of total reads for each individual sample that was attributable to blood-, semen-, saliva-, vaginal secretions-, menstrual blood- and skin-specific markers. Agglomerative hierarchical clustering analysis was also performed.

Multiplexed targeted mRNA NGS assay for body fluid identification - Ion S5 platform

The Ion S5 (ThermoFisher, Applied Biosystems) was performed according to manufacturer's instructions using either a manual library preparation protocol (MAN0007450, Revision A.0) or the automated Ion Chef protocols. To prepare the libraries, 50 ng RNA and 30 amplification cycles were used. Pooling and diluting of the libraries was recommended as the following: Ion 314 chip – pooling to 25 pM, end concentration of 4 pM; Ion 316/218 chips – pooling to 100 pM, end concentration of 16 pM; Ion 520/530 chips – pooling to 100 pM, end concentration of 6.25 pM. 500 flows were used for sequencing. The Ion AmpliSeq^{TM} RNA plugin was used to provide coverage summaries and read counts per amplicon.

Interpretation of mRNA sequencing data for body fluid identification – Statistical Methods

The NGS sequencing data was used to build a probabilistic model that predicts that origin of a stain. Our approach uses partial least squares followed by linear discriminant analysis to classify samples into six commonly occurring forensic body fluids. The model incorporates quantitative information (NGS read counts) rather than just presence/absence of markers. It allows for visualization of important markers and their correlation with the different body fluids.

IV. Findings

**detailed results are provided in the 5 publications (Appendix A) resulting from this work*

A. Multiplexed targeted mRNA NGS assay for body fluid identification

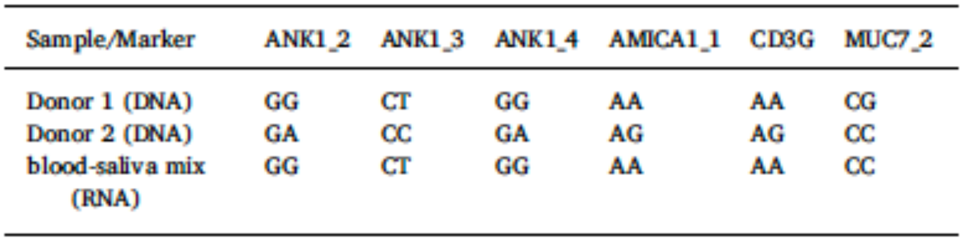
Two multiplexed targeted mRNA NGS assays for body fluid identification were developed, one using the Illumina MiSeq platform and one using the Ion PGM/S5 platforms (Thermofisher, Applied Biosystems). The MiSeq assays is a fully optimized and finalized 33plex system containing 6 biomarkers for blood, 6 biomarkers for semen, 6 biomarkers for saliva, 4 biomarkers for vaginal secretions, 5 biomarkers for menstrual blood and 6 biomarkers for skin. The Ion PGM/S5 assay is a prototype assay and is currently still under development and further optimization. It is a 29plex system containing 4 biomarkers for blood, 6 biomarkers for semen, 6 biomarkers for saliva, 4 biomarkers for vaginal secretions, 4 biomarkers for menstrual blood and 5 biomarkers for skin. Both assays result in a definitive identification of each of the target body fluids/tissues.

A full validation and optimization of the MiSeq 33plex assay was performed. All included biomarkers demonstrated a high degree of specificity to their target body fluid/tissue with little to no cross-reactivity with non-target body fluids. Initial performance checks were performed including: 1) biomarker input sensitivity (50 ng total RNA determined to be optimal with some body fluids detectable with only 5-10 ng total RNA), admixed body fluid samples (assay permits identification of multiple components in admixed samples, although additional work is needed to determine appropriate thresholds for the accurate identification of minor component fluids), repeatability (good when the raw counts were normalized as a percent of the total reads and apportioned into the appropriate body fluid specific classes), species specificity of blood biomarkers (SPTB and ALAS2 possibly not exclusive to humans and primates, but overall pattern expression could be used to differentiate humans and primates) and organ tissue specificity (majority of biomarkers had no significant cross reactivity with organ tissue samples). A blind study was also performed in which the origin of main body fluid components were successfully identified. Additionally, a novel probabilistic method was developed based on PLS-DA for prediction of the origin of a stain. For the first time, quantitative information (NGS) read counts have been incorporated into a model, which on our data performed better than a model that includes only presence/absence. Our method also performed better than previously suggested methods that have the option to include quantitative data.

The MiSeq and the prototype Ion NGS body fluid assays were used in a collaborative EUROFORGEN/EDNAP exercise in order to test the efficacy of targeted mRNA sequencing to identify body fluids. This collaborative study included participation from 17 laboratories world-wide. The results demonstrated moderate to high count values in the body fluid or tissue of interest with little to no counts in non-target body fluids. There was some inter-laboratory variability in read counts, but overall the results of the laboratories were comparable in that highly expressed markers showed high read counts and less expressed markers showed lower counts. We performed a partial least squares (PLS) analysis on the data, where blood, menstrual blood, saliva and semen markers and samples clustered well. The results of the collaborative exercise support the use of targeted mRNA sequencing as a reliable body fluid identification method that could be added to the repertoire of forensic NGS panels.

B. cSNP assay

In the above described assays, we demonstrated the successful use of a targeted multiplex RNA NGS assay for body fluid identification. However, a goal of the current work was not only to provide an assay for body fluid identification but to also permit an association of an individual body fluid with a particular donor in admixed samples. Such an association would not be possible with any other current body fluid identification methods and is unique to mRNA profiling for body fluid identification. This association was accomplished through the use of coding region SNPs (cSNPs) within individual mRNA targets. Within tissue specific sequences, we looked for coding region variants that discriminate European individuals the most. We successfully tested a prototype targeted NGS assay using the Illumina MiSeq platform in 188 European individuals for the detection of the 35 selected cSNPs on a genomic DNA level. The corresponding RNA cSNP panel showed good specificity for blood, semen and menstrual blood. For saliva, vaginal secretions and skin, the marker design needs to be optimized with special attention to cross reactivity with DNA contaminants. In a proof-of-principle experiment, we demonstrated that with our targeted DNA and RNA NGS assays, we were able to assign a body fluid to a specific individual. For this experiment, we analyzed a blood-saliva mixture. Using reference samples from the donors in the mixture, we generated reference cSNP genotypes using the gDNA cSNP assay. When the admixed sample was analyzed with the RNA cSNP assay, the genotypes at each of the included blood and saliva biomarkers were analyzed. The most discriminating cSNPs are provided in the table below.



Sample/Marker	ANK1_2	ANK1_3	ANK1_4	AMICA1_1	CD3G	MUC7_2
Donor 1 (DNA)	GG	CT	GG	AA	AA	CG
Donor 2 (DNA)	GA	CC	GA	AG	AG	CC
blood-saliva mix (RNA)	GG	CT	GG	AA	AA	CC

As can be seen from these results, the blood in the admixed sample could only have originated from donor 1 and the saliva sample originating from donor 2. Therefore, for the first time we were able to successfully assign the body fluids in an admixtures to the respective donors on the basis of RNA coding SNPs in body fluid specific transcripts. An additional manuscript describing the development and initial testing of the prototype cSNP assay is currently being prepared and will be submitted for publication.

Despite the success of this initial assay, more extensive studies including optimized assay development need to be performed. For the MiSeq platform, the initial prototype assay was designed using existing 'off-the-shelf' assays. Assays were not available for all cSNPs of interest and there is not a design feature in order to create custom primer sets. Therefore, we are now currently focusing efforts on the Ion S5 platform in order to design a custom-built cSNP assay that will permit the incorporation of additional cSNP markers for each body fluid. We are hopeful that this will improve the specificity and discriminatory capacity of the cSNP assay. This assay will be designed so that a simultaneous identification of the body fluid of origin and association of body fluid to donor in admixed samples can be performed. This work is currently underway.

V. Implications for criminal justice policy and practice in the United States

The recovery of a DNA profile from the perpetrator in criminal investigations provides valuable ‘source level’ information for investigators. However, a DNA profile does not reveal the circumstances by which biological material was transferred. This contextual information (or ‘activity level’) can be obtained by a determination of the tissue or fluid source of origin of the biological material as it indicates some behavioral activity on behalf of the individuals that resulted in its transfer from the body. This project sought to, and successfully improved upon, an established molecular based method for body fluid identification, namely mRNA profiling, by the development of a targeted multiplexed next generation sequencing assay. The next generation sequencing assays developed here provide not only probative ‘activity level’ information in criminal investigations, but uniquely provide novel probative information, coding region SNPs (or RNA-SNPs). RNA-SNPs will, for the first time, permit an association of a DNA profile to a specific body fluid or tissue in admixed samples. This source attribution of biological material cannot be obtained using any other body fluid identification method (e.g. CE or real time PCR analysis, epigenetics, proteomics). The assays develop here could also result in the development of a commercial product for facile transfer to operational crime laboratories.

Appendix A. List of Publications

[1] Hanson E.K., Ingold S., Haas C. and Ballantyne J. Messenger RNA Biomarker Signatures for Forensic Body Fluid Identification Revealed by Targeted RNA Sequencing. Forensic Science International Genetics 34 (2018), p. 206-221.

[2] Dorum G., Ingold S., Hanson E., Ballantyne J., Snipen L. and Haas C. Predicting the Origin of Stains from Next Generation Sequencing mRNA Data. Forensic Science International Genetics. 34 (2018), p. 37-48.

[3] Ingold S., Dorum G., Hanson E., Berti A., Branicki W., Brito P., Elsmore P., Hetting, K.B., Giangasparo F., Gross T., Hansen S., Hanssen E., Kampmann M., Kayser M., Laurent F., Morling N., Mosquera-Miguel A., Parson W., Phillips C., Porto M., Pospiech E., Roeder A., Schneider P., Schulze Johann K., Steffen, C.R., Syndercombe-Court D., Trautmann M., van den Berge M., van der Gaag K., Vannier J., Verdoliva V., Vidake A., Xavier C., Ballantyne J. and Haas C. Body Fluid Identification Using a targeted mRNA Massively Parallel Sequencing Approach – Results of a EUROFORGEN/EDNAP Collaborative Exercise. Forensic Science International Genetics. 34 (2018), p. 105-115.

[4] Hanson, E., Ingold, S., Haas, C. and Ballantyne, J. Targeted multiplex next generation RNA sequencing for tissue source determination of forensic samples.
Forensic Science International: Genetics Supplement Series (2015), http://dx.doi.org/10.1016/j.fsigss.2015.09.175

[5] Ingold S., Haas C., Dorum G., Hanson E. and Ballantyne J. Association of a Body Fluid with a DNA Profile by Targeted RNA/DNA Deep Sequencing. Forensic Science International Genetics: Supplement Series 6 (2017), e112-e113.